— Supplementary Material — Fill in Fabrics: Body-Aware Self-Supervised Inpainting for Image-Based Virtual Try-On

Hasib Zunair[1]
hasibzunair@gmail.com
Yan Gobeil[2]
yan.gobeil@decathlon.com
Samuel Mercier[2]
samuel.mercier@decathlon.com
A. Ben Hamza[1]
hamza@ciise.concordia.ca

[1] Concordia University
Montreal, QC, Canada
[2] Decathlon Canada
Montreal, QC, Canada

This supplementary material includes the implementation details of our method, additional experimental results, and ablation studies.

1 Implementation Details

Data Preprocessing. Given a reference person image $\mathbf{I}$, we first compute the mask $\mathbf{M}$ using a human parser [10]. We also predict the 18-keypoint pose heatmap $\mathbf{M}_{\text{pose}}$ using out-of-the-box 2D pose estimator [1, 2]. All images and masks are of resolution 256×192, and normalized to have values in $[-1, 1]$.

Architecture. Our Fill in Fabrics (FIFA) model consists of a Fabricator, Segmenter, Warper and Fuser. The Segmenter consists of conditional generative adversarial networks (CGANs) where the generators G_c is a U-Net model [13] and G_p is a Residual U-Net [18]. The discriminators are pix2pixHD [15]. The Warper consists of a Spatial Transformer Network (STN) [7] with Thin Plate Splines (TPS) [9], and an additional refinement network, which is a U-Net model. STN is comprised of five convolutional layers and a max-polling layer with a stride set to 2. The Fuser also has a CGAN G_m, which is a U-Net model.

Model Training. FIFA is trained in two steps. First, we train the Fabricator on the clothing images only by masking the clothing images; thereby producing input-output pairs. After training the Fabricator, the weights of the encoder-decoder network $\mathcal{F}_s$ are used for initialization when training the Warper, which we refer to as masked cloth modeling (MCM). We train the Segmenter, Warper (i.e. with weights pre-trained from Fabricator) and Fuser for 100 epochs, which takes roughly seven days on a single NVIDIA RTX 3080 GPU. During training, the target clothing items are the same as the clothing in the reference person image since it is not possible to acquire triplets to compute loss with respect to ground truth. We

use Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$, an initial learning rate of 0.0002, and a batch size equal to 1. A batch size larger than 1 gives a GPU memory error.

Model Testing. After training, we discard the Fabricator. For computing the evaluation metrics, we follow the same setting during training (i.e. target clothing is same as in the reference person image). For more realistic testing, the target clothing image is different from the one in the reference person image. We also test on easy, medium and hard cases to get a better understanding of the try-on quality by the different methods. In addition, we provide generalization tests with the aim to train on one dataset and test on a completely different dataset for the same virtual try-on task.

2 Additional Experimental Results

Qualitative Comparison Results:

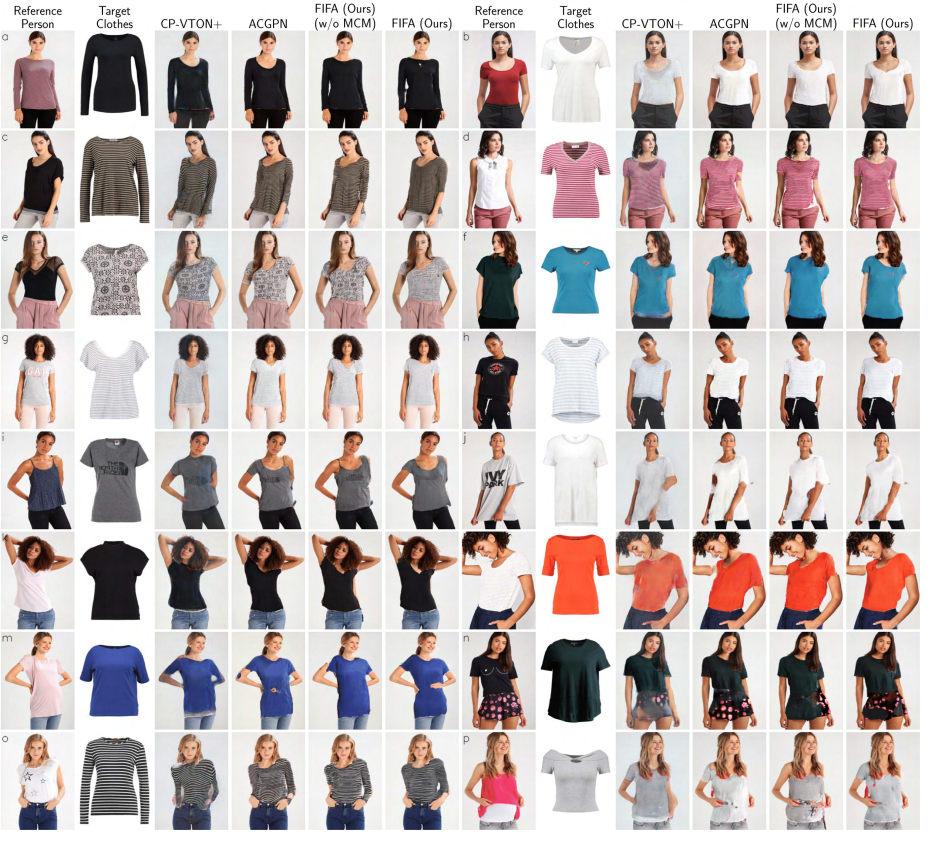

Figure 1: Visual visual results of image-based try-on outputs. FIFA generates realistic try-on results, which preserve the texture and embroidery of the target clothing and handle better the complex poses.

Figure 2: Visual results of image-based try-on outputs. FIFA generates realistic skin parts, completely removes old clothing, differentiate front and back part of clothing and produces consistent skin color during skin synthesis

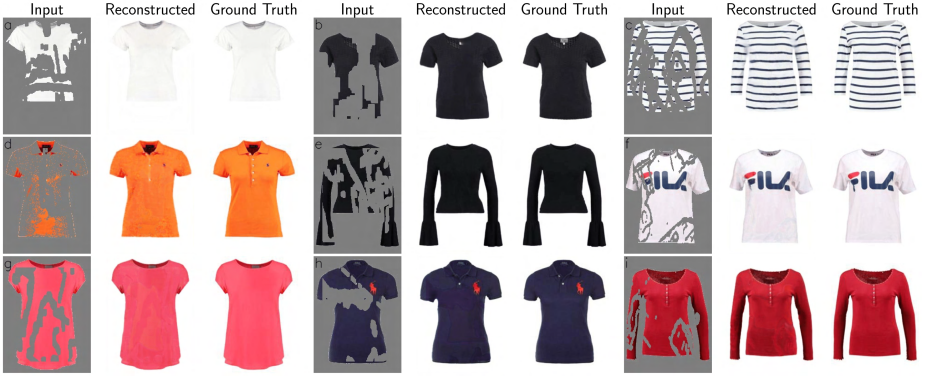

Figure 3: Masked clothing images and their corresponding reconstructed clothing outputs from Fabricator. The Fabricator model is able to capture the overall structure of the clothing item (e.g., both full and short sleeves) when a partial input is provided.

Effectiveness of Masked Cloth Modeling:

Figure 4: Visual results demonstrating the importance of the MCM objective. MCM helps preserve and/or synthesize body parts in complex poses (i.e. body-aware) as well as accurately warp the target clothing.

Effectiveness of Residual Blocks:

Figure 5: Ablation study of segmentation results demonstrating the effectiveness of residual blocks in Segmenter. Segmenter with RBs helps predict better the fine-grained semantic layout of body parts.

Ablation Study on Residual Blocks:

Table 1: Ablation study of using residual blocks (RBs) in G_p compared to G_p, G_c and G_m. Bolface numbers indicate better performance.

	SSIM (↑)				
Method	VITON	VITON-E	VITON-M	VITON-H	**FID (↓)**
w/o RBs	0.848	0.857	0.850	0.830	16.41
RBs (G_p)	**0.886**	**0.890**	**0.880**	**0.865**	**13.46**
RBs (G_p, G_c, G_m)	0.864	0.874	0.859	0.844	15.43

Table 2 (first two rows) further corroborates our findings in the sense that improvements are observed in terms of both SSIM and FID. Unlike the SSIM metric, FID does not capture the sharp high-frequency details. We can also see substantial improvements across across all, easy, medium and hard cases. The claim that MSC exploits global context is also verified in Table 2 (middle two rows), which shows a significant improvement in terms of the FID metric that captures distributional similarity. The SSIM scores (easy, medium, hard and aver-

age) are also consistently improved when using MSC. In addition, our approach outperforms DCTON, which uses an adversarial loss and an additional cycle consistency loss term, when using MSC (i.e. FIFA with MSC). Table 2 (bottom two rows) demonstrates the benefit of residual blocks in improving the SSIM and FID scores. Residual blocks are able to retain the fine-grained features from the initial features of the input that are lost during the downsampling stage in the encoder, and hence they cannot be fully recovered during the upsampling stage in the decoder, leading to poor quality of the segmentation maps and hence poor try-on results. This is largely attributed to the limited capability of the U-Net architecture used in most of the existing methods. Residual blocks are proven to have better information propagation and effectively learn effective representations of the input data, especially for image recognition tasks [6].

Table 2: Ablation study of the different modules (MCM, MSC and RBs) on VITON, VITON-E, VITON-M and VITON-H test sets. Boldface numbers indicate better performance.

	SSIM (↑)				
Method	VITON	VITON-E	VITON-M	VITON-H	**FID** (↓)
w/ MCM	**0.886**	**0.890**	**0.880**	**0.865**	**13.46**
w/o MCM	0.868	0.879	0.865	0.846	13.65
w/ MSC	**0.868**	**0.879**	**0.865**	**0.846**	**13.65**
w/o MSC	0.853	0.861	0.852	0.832	15.59
w/ RBs	**0.853**	**0.861**	**0.852**	**0.832**	**15.59**
w/o RBs	0.848	0.857	0.850	0.830	16.41

Generalization to In-The-Wild Virtual Try-On:

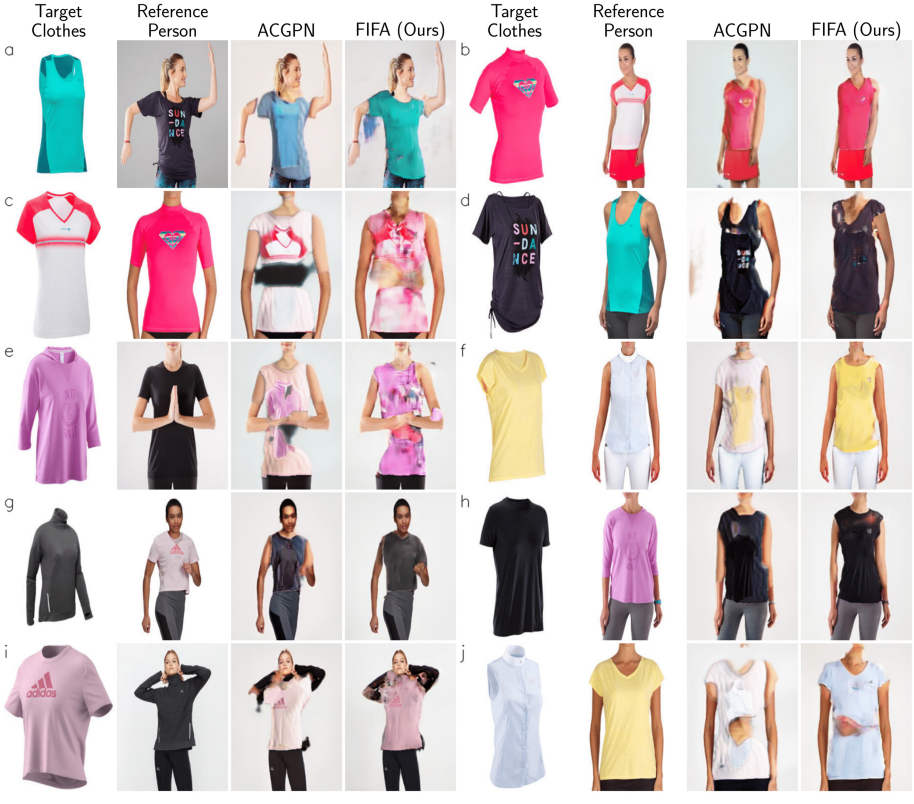

Figure 6: Qualitative comparison of try-on outputs of FIFA against ACGPN demonstrating the generalization to in-the-wild virtual try-on.

Table 3: Robustness tests of models trained on VITON and tested on the DecaWVTON dataset. Our FIFA model outperforms the strongest baseline in terms of both metrics.

Method	**SSIM** (↑)	**FID** (↓)
ACGPN [16]	0.927	152.43
FIFA	**0.952**	**147.62**

Figure 7: Qualitative comparison of warped clothing outputs of FIFA against ACGPN demonstrating the generalization to in-the-wild virtual try-on.

Figure 8: Qualitative comparison of try-on outputs of test images of the clothing item and reference person from the VITON test set and DecaWVTON, respectively.

3 Diversity

The vast majority of images in the VITON dataset are of white female models, and there are only very few dark-skinned models. Interestingly, our method is able to preserve the skin color of under-represented models and performs better than the baselines CPVTON+ [11] and ACGPN [16]. We believe that this better performance is largely attributed to the better warping of target clothing, resulting in an overall better human synthesis. Figure 9 shows some visual examples.

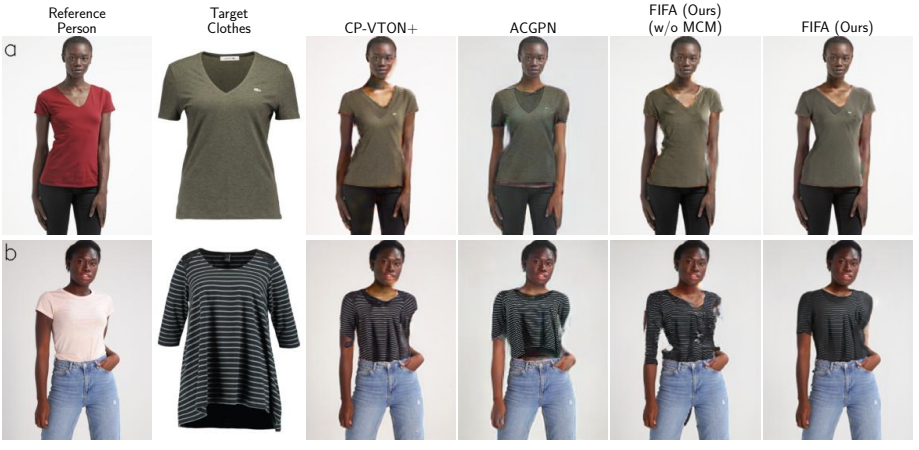

Figure 9: Comparison of virtual try-on outputs against CPVTON+ [11] and ACGPN [16] in cases of under-represented models. Our method can accurately synthesize skin color and outperforms the baseline methods.

References

[1] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[3] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. VITON-HD: High-resolution virtual try-on via misalignment-aware normalization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 14131–14140, 2021.

[4] Jean Duchon. Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In *Constructive Theory of Functions of Several Variables*, pages 85–100. Springer, 1977.

[5] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. VITON: An image-based virtual try-on network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 7543–7552, 2018.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[7] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, 2015.

[8] Surgan Jandial, Ayush Chopra, Kumar Ayush, Mayur Hemani, Balaji Krishnamurthy, and Abhijeet Halwai. SieveNet: A unified framework for robust image-based virtual try-on. In *Proc. IEEE Winter Conference on Applications of Computer Vision*, pages 2182–2190, 2020.

[9] Nikolay Jetchev and Urs Bergmann. The conditional analogy GAN: Swapping fashion articles on people images. In *Proc. IEEE International Conference on Computer Vision Workshops*, pages 2287–2292, 2017.

[10] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[11] Matiur Rahman Minar, Thai Thanh Tuan, Heejune Ahn, Paul Rosin, and Yu-Kun Lai. CP-VTON+: Clothing shape and texture preserving image-based virtual try-on. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[12] Bin Ren, Hao Tang, Fanyang Meng, Runwei Ding, Ling Shao, Philip HS Torr, and Nicu Sebe. Cloth interactive transformer for virtual try-on. *arXiv preprint arXiv:2104.05519*, 2021.

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[14] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proc. European Conference on Computer Vision*, pages 589–604, 2018.

[15] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018.

[16] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 7850–7859, 2020.

[17] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. VTNFP: An image-based virtual try-on network with body and clothing feature preservation. In *Proc. IEEE International Conference on Computer Vision*, pages 10511–10520, 2019.

[18] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual U-Net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018.